Flower Classification and Identification: Under Data Mining Tasks

ZHANG CHENYANG, 57126363, City University of Hong Kong

SUN QIXUAN, 57125250, City University of Hong Kong

	Topic	Model Construction	Coding	PowerPoint	Presentation
ZHANG	√	√	√	√	√
SUN	√	√	√	√	√

Abstraction

PictureThis is a mobile app that uses artificial intelligence to identify plants. Users only need to upload a photo of a plant, and after a while the APP will be able to determine what kind of plant it is. Image recognition is a popular field in AI. In this study, we aim to realize the function of flower identification, by accomplish some data mining tasks, such as clustering and hashing methods, to extract the implicit pattern of flowers. The dataset is from Kaggle, which contains 4572 photographs of different kinds of flowers. The quality of these tasks is evaluated by evaluation matrices. The result shows a quite strong ability to identify flowers.

1. Introduction

Plant classification and recognition has been a popular research topic in the fields of computer vision and machine learning. With the continuous enrichment of biological diversity and the demand for applications in agriculture and other fields, accurately and quickly identifying flower species has important scientific value and broad application prospects. Traditional flower identification relies on the expertise of botanists, which is time-consuming and costly. With the development of deep learning technology, using computers to automatically perform image recognition has become an effective method to solve this problem. Current mainstream flower recognition technologies such as convolutional neural networks (CNN) based on deep learning have proven to be particularly effective in processing complex image data. This research aims to explore and optimize deep learning models for flower image classification, by applying some data mining tasks, including clustering, convolutional neural networks, hashing, data enhancement and semi-supervised learning. Through experimental verification, we will evaluate the performance of our model in the flower recognition task and explore how to improve the accuracy of the model in complex natural scenes, and finally present the results and discuss their scientific and practical application implications.

2. Background

2.1 Importance of research

Automated flower identification systems can help professionals and amateurs quickly identify plant species, promoting botanical research and biodiversity conservation. In addition, this technology also shows broad application potential in areas such as nature reserve management and education. There are many apps on the market that automatically identify plants, such as PictureThis (shown in Fig. 1).

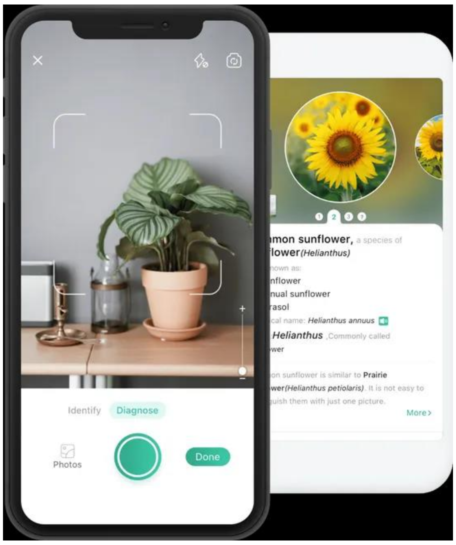

mon sunflower, a species of
lower (Helianthus)
nown as:
nflower
nual sunflower
rasol
cal name: Helianthus annuus
Helianthus .Commonly called
wer
on sunflower is similar to Prairie
wer(Helianthus petiolaris). It is not easy to
guish them with just one picture.
More ›
Identify
Diagnose
Photos
Done

Figure 1. Plants identification in PictureThis

2.2 Challenge

Although related technologies have made significant progress in recent years, flower image recognition still faces a number of technical challenges, including but not limited to the quality and diversity of images, changes in flowers under different environmental conditions, and subtle differences between similar species. In addition, how to automatically extract effective features from large-scale unlabeled flower image datasets and perform accurate classification is also the focus of this study.

2.3 Dataset

The dataset includes 9 categories in the dataset, and each category contains 408 images. The dataset is split into a labeled training set, an unlabeled training set, and a validation set, with the ratio of 78.44%, 1.96%, 19.6%. There are in total 4572 pieces of photographs in total and the dataset size is about 4.26GB.

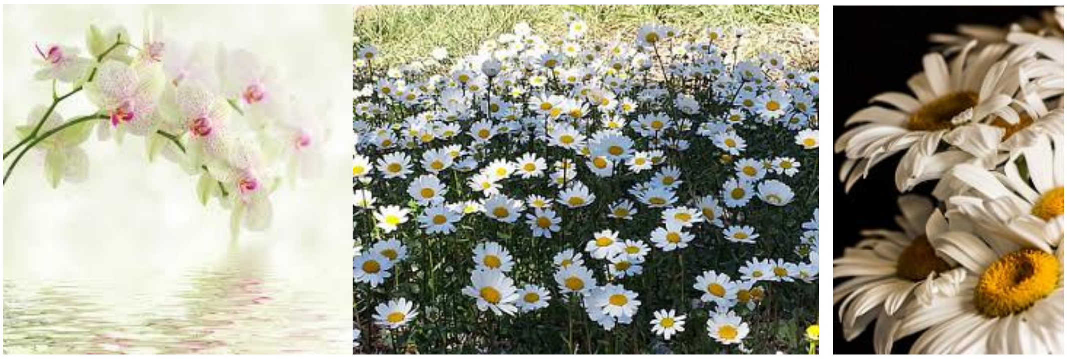


Fig. 2. Image data in the dataset

3. Methodology

3.1 Data Augmentation

Data augmentation is a technique that is especially useful when working with image or audio data. This technique generates new and varied data points by applying a series of random transformations to existing data, aiming to increase the diversity of a dataset without actually collecting more data. In this section, we also use data augmentation to preprocess our data. By applying rotations, scaling and flips to the raw data, visually different instances of the data are generated. These variations increase the diversity of the dataset and help the model learn important features when similar objects are viewed from different angles and under different size conditions.

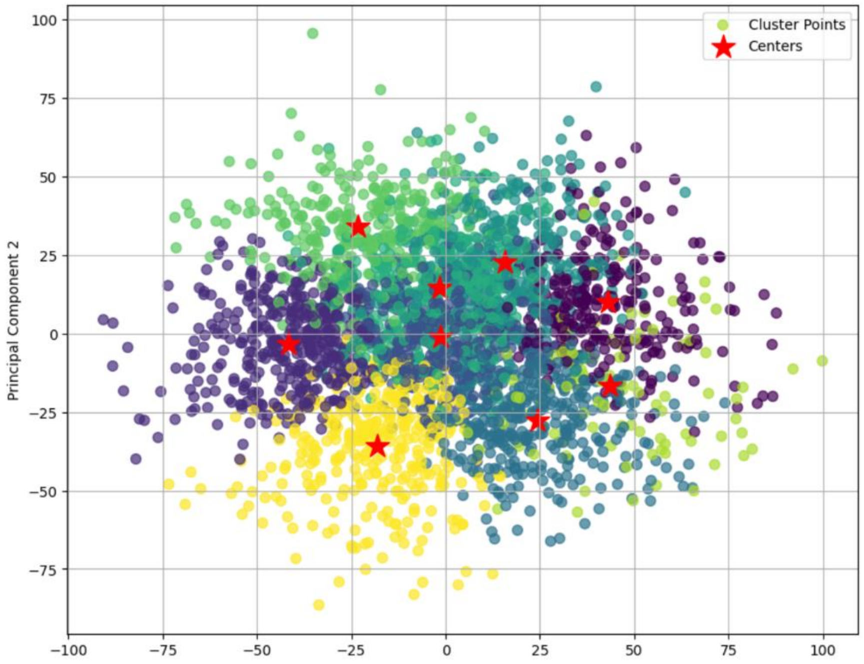


Fig 3. Clusters by K-means

3.2 Semi-supervised Learning

There may be a large number of unlabeled flower images in the dataset. Through semi-supervised learning methods, these unlabeled data can be used to enhance the learning effect of the model. This method can improve the generalization ability of the model with limited labeled data.

3.2.1 Fine-tuning strategy

Fine-tune the model according to the characteristics of the clustering. The learning rate, optimization algorithm, or other hyperparameters can be adjusted independently for each cluster to adapt to the characteristics of different clusters.

3.2.2 Identify and handle anomalies

Use clustering results to identify abnormal data or outliers, and then decide whether to delete the data or perform special processing (such as special annotation or separate training).

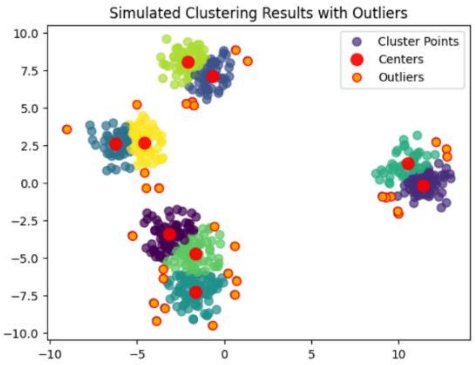


Fig 4. Clustering Results with Outliers

3.3 Clustering

Clustering can help discover natural groupings that exist in our data set, revealing similarities and differences between different flower categories. This understanding helps to adjust data preprocessing and augmentation strategies to ensure that they are more suitable for the actual data distribution. Here clustering is used in both data augmentation and semi-supervised learning processes.

After clustering, the data augmentation method can be adjusted based on the characteristics of each cluster. For example, if a cluster contains many dark flower pictures, brightness adjustment can be added to these pictures; if the pictures of a certain cluster mainly include flowers from a specific angle, data augmentation of angle transformation can be added. Clustering these images that are suitable for the same adjustment can make the features of the cluster more unified and obvious.

K-means

In the flower recognition project, the k-means clustering algorithm is primarily employed to organize and categorize large sets of flower images efficiently. k-means helps in clustering flower images that have similar features such as color, shape, or direction. It can also be used to identify and create centroids that summarize the features of various clusters. These centroids can represent common patterns in the data, which can be useful for feature reduction and simplification of the problem space. Considering the dataset, k-means is particularly well-suited for handling large datasets due to its relatively simple implementation and fast processing times.

Time complexity is $O(\boldsymbol{t}\times\boldsymbol{k}\times\boldsymbol{n}\times\boldsymbol{f})$, where t is the iteration times, k is the number of clusters and here $k = 9$, n is the number of samples, and f is dimensionality of features. Overall, k-means clustering provides a robust method for organizing and simplifying flower image datasets, enhancing the effectiveness of subsequent analysis and recognition tasks in the flower recognition project.

Hierarchical Clustering

Hierarchical clustering can be used for image segmentation to contribute to identify and extract features in images to benefit the following image identification process. Different with K-means, it does not need to specify the number of clusters before clustering, but has high computational complexity, especially when dealing with large data sets.

Generated dendrogram represents the similarity relationship between data points and the hierarchical structure of its clustering and is often used to visually show how data points are combined into larger clusters step by step.

In flower identification tasks, dendrogram can help identify which categories of images are close or more different to each other in features. This information can be used to optimize the parameters of the clustering algorithm to better distinguish images that may be visually similar but actually belong to different categories.

Hierarchical clustering algorithms typically have a complexity of $O(n^3)$ when using basic implementations, as it requires recalculating distances between clusters for each merge operation. But by employing a min-heap data structure to maintain a priority queue of cluster distances in our model, the complexity can be reduced to $O(n^2 \log n)$.

3.4 Locality Sensitive Hashing (LSH)

The application of locality-sensitive hashing in the flower recognition project mainly focuses on optimizing the data processing process, improving system response speed and the ability to process large-scale image data sets. Locality-sensitive hashing is often used to find similar data items quickly, which is very useful in flower recognition. When performing cluster analysis of flower images, LSH can be used to preprocess the data, effectively identify and remove duplicates, and quickly estimate the similarity between images. This can help clustering algorithms organize data more efficiently, significantly improving the speed and quality of clustering, especially when dealing with large-scale data sets.

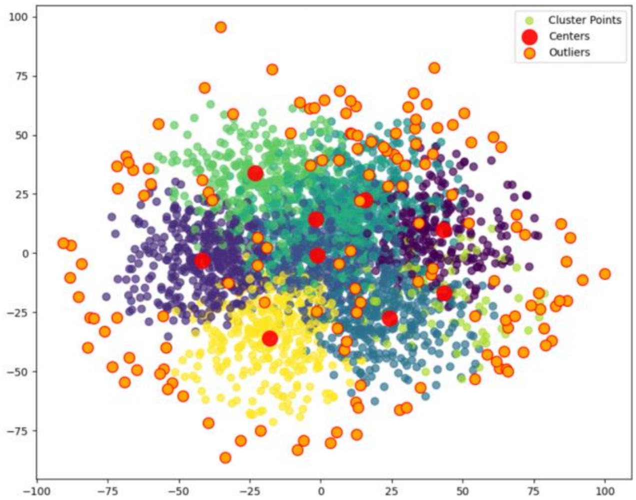


Fig. 5. Clustering after LSH

3.5 Convolutional Neural Network

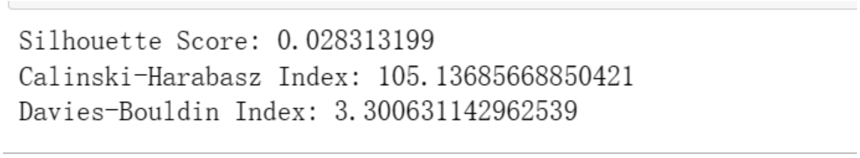
This section defines a model based on a convolutional neural network (CNN). It integrates the Squeeze-and-Excitation (SE) module to improve the network's ability to express image features. This network structure is widely used in image recognition and classification tasks, and its design concept and structural characteristics make it perform well when processing image data. The following is a detailed introduction to this network structure and its application in data mining:

- **Convolutional Layers**: These are the building blocks of our model, extracting detailed textures and local features from images, empowered by nonlinear activation functions like ReLU to handle complex patterns.
- **Batch Normalization Layers**: These layers are essential for accelerating the training process, ensuring model stability, and mitigating issues like internal covariate shift by normalizing layer inputs.
- **Squeeze-and-Excitation Module**: At the heart of our enhanced performance is the SE module, which fine-tunes the interdependencies between channels, enabling the model to focus more accurately on relevant features through adaptive recalibration.
- **Pooling Layers**: By reducing dimensionality and enhancing the model's generalization capabilities, these layers ensure robustness and positional invariance of features.
- **Adaptive Average Pooling**: Following up, this layer condenses the feature map to a manageable size, streamlining the network for the classification stage.
- **Fully Connected Layers**: These layers decipher high-level feature relationships to deliver final classification outcomes, with dropout layers included to curb overfitting.

4. Result & Conclusion

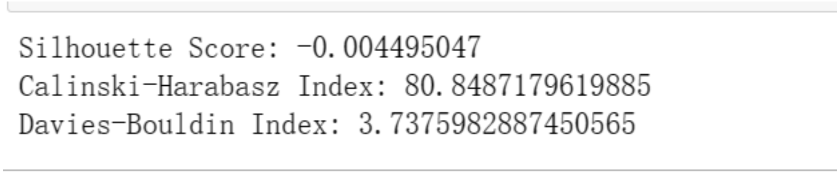
4.1 Evaluation of Data Mining Tasks

The accuracy of the model has been significantly improved after data augmentation and semi-supervised learning. We use indicators such as silhouette score to measure the improvement brought by data mining tasks.



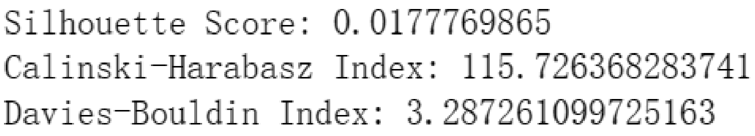
```
Silhouette Score: 0.028313199
Calinski-Harabasz Index: 105.13685668850421
Davies-Bouldin Index: 3.300631142962539
```

Indicators of data augmentation by K-means



```
Silhouette Score: -0.004495047
Calinski-Harabasz Index: 80.8487179619885
Davies-Bouldin Index: 3.7375982887450565
```

Indicators of data augmentation by hierarchical clustering

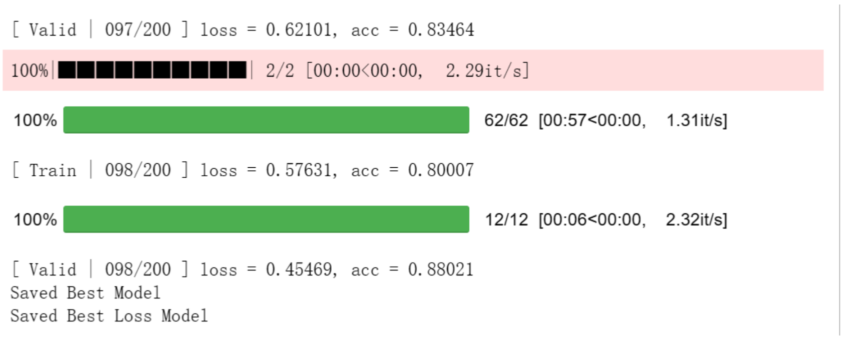


```
Silhouette Score: 0.0177769865
Calinski-Harabasz Index: 115.726368283741
Davies-Bouldin Index: 3.287261099725163
```

Indicators of data augmentation by LSH

4.2 Result

This project implemented various data mining tasks including data augmentation, semi-supervised learning, k-means, hierarchical clustering, locality-sensitive hashing (LSH), and convolutional neural networks (CNNs) to accomplish the goal of flower identification based on the dataset we chose. Through rigorous experimentation and validation, the final model recognition accuracy reached a maximum of 0.88, and the loss of the verification set reached 0.45. Compared with before the clustering task, the accuracy increased by 10%, achieving an impressive result. We believe that more subsequent training will make the model more intelligent and have stronger capabilities in classification.



```
[ Valid | 097/200 ] loss = 0.62101, acc = 0.83464
100%|██████████| 2/2 [00:00<00:00,  2.29it/s]
100%                               62/62  [00:57<00:00,   1.31it/s]
[ Train | 098/200 ] loss = 0.57631, acc = 0.80007
100%                               12/12  [00:06<00:00,   2.32it/s]
[ Valid | 098/200 ] loss = 0.45469, acc = 0.88021
Saved Best Model
Saved Best Loss Model
```

Fig. 6. Best model and best loss model

The application of clustering algorithms such as k-means and hierarchical clustering provided insights into the underlying structures within the flower image dataset, enabling efficient grouping and organization of similar images. Furthermore, the utilization of locality-sensitive hashing expedited the process of identifying similar images, facilitating rapid retrieval and matching of flower images. These tasks enable the model to generate clearer categories, then instruct the better performance of data augmentation and semi-supervised learning, at last benefit the function of identification of the model.

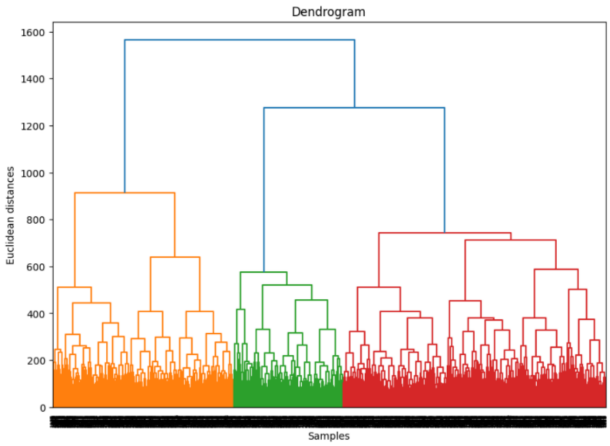


Fig. 7. Dendrogram generated by Hierarchical Clustering

Dendrogram represents the similarity relationship between data points and the hierarchical structure of its clustering and is used to visually show how data points are combined into larger clusters step by step. The horizontal line represents the merge operation, and the two vertical lines or groups connected by each horizontal line indicate that the two groups are merged into a new group at that level. The vertical line represents the history of the group, i.e. from which previous mergers the group was formed. Longer vertical lines indicate that the group existed for a longer period before merging. The height of the horizontal line represents the distance or dissimilarity when merging groups. The larger the height, the greater the difference in characteristics between the two merged groups. Therefore, height can be considered as an indicator of differences between groups. The horizontal line at the bottom represents the merge with the smallest distance or similarity, meaning these data points are very similar to each other. As the observations move upward, the merged groups become larger and represent greater dissimilarity.

Also, we utilize clustering methods to visualize a representative image of each cluster in a dataset in a matrix layout.

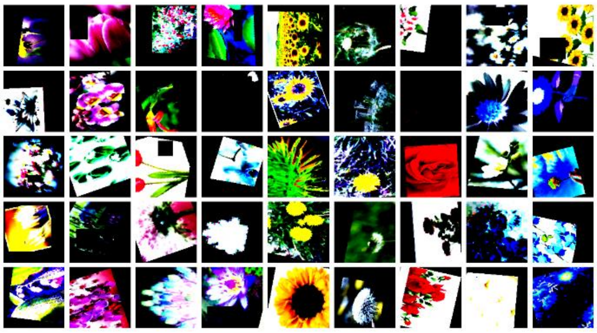


Fig 8. Standard Conversion Model

The integration of data augmentation techniques allowed for the generation of diverse and augmented datasets, through adjustments such as rotation and scaling, enhancing the robustness of the model against variations in different environmental conditions, orientations, and photography conditions of flowers. Additionally, the incorporation of semi-supervised learning methodologies facilitated the utilization of large amount of unlabeled data in the dataset to improve model generalization and performance.

The pivotal component of the system was the convolutional neural network, which served as the backbone for learning discriminative features from flower images. Leveraging the power of deep learning, the model exhibited exceptional capability in distinguishing between different flower species with high accuracy after been trained 300 times.

4.3 Conclusion

In conclusion, the flower recognition system presented in this study achieved remarkable success in accurately identifying flower species, attaining an impressive accuracy rate of 87%. By leveraging a combination of data augmentation, semi-supervised learning, clustering techniques, locality-sensitive hashing, and convolutional neural networks, the system demonstrated robustness and effectiveness in handling the complexities inherent in flower image datasets.

The integration of data augmentation techniques contributed to the augmentation of datasets, enhancing the model's resilience to variations in environmental conditions. Moreover, the adoption of semi-supervised learning methodologies enabled the utilization of unlabeled data, augmenting the model's learning capacity and improving its performance.

Clustering algorithms such as k-means and hierarchical clustering provided valuable insights into the underlying structures of the dataset, facilitating efficient organization, and grouping of similar images. The incorporation of locality-sensitive hashing sped up image similarity retrieval, enabling rapid identification and matching of flower images.

Central to the success of the system was the convolutional neural network, which effectively learned discriminative features from flower images, enabling accurate species classification. Moving forward, the integration of these techniques holds promise for advancing the field of flower recognition, with implications for various applications including botany, agriculture, and environmental conservation.

The integration of multiple methodologies in this study demonstrates a successful approach to tackling the challenges of flower recognition in diverse and uncontrolled environments. The high accuracy achieved underscores the efficacy of combining data augmentation, semi-supervised learning, advanced clustering techniques, locality-sensitive hashing, and deep learning in a cohesive workflow.

Our results suggest that such a multifaceted approach can significantly enhance the performance of image recognition systems in botany and related fields. This study not only advances the state of flower recognition technology but also sets a precedent for future research in automated plant classification systems. By continuing to refine these techniques and expand our dataset, further improvements in accuracy and system robustness are expected.

Moreover, the methodologies developed and tested in this project have the potential for broader applications, such as in ecological monitoring, agricultural automation, and biodiversity research, where accurate and efficient plant recognition can play a crucial role in conservation and management strategies.

This project serves as a model for future explorations into plant recognition, providing a scalable and efficient solution capable of adapting to the growing needs of environmental and biological sciences.

5 Discussion

Semi-supervised learning improved learning efficiency by utilizing abundant, often unused unlabeled data, suggesting potential for bridging supervised and unsupervised learning gaps. Locality-sensitive hashing enhanced the speed of image retrieval but depends heavily on parameter tuning, such as hash tables and key length. The CNN was crucial for high classification accuracy, adept at extracting complex features from flower images. Future directions could include integrating attention mechanisms or explainable AI to improve transparency and interpretability.

Future Work

To further optimize the functionality and efficiency of your flower identification project, more data mining tasks can be implemented. Improvements in feature engineering, such as edge detection and texture analysis, can significantly improve the performance of clustering and classification algorithms. We can also Implement interactive or active learning strategies that enable the system to learn from user feedback and continuously optimize itself. For example, users can provide correct information when recognition results are inaccurate, and the system updates its learning accordingly. Moreover, experiment with different network architectures or hyperparameter tuning may find a better model configuration for the flower recognition task. These optimization measures help build a more powerful, flexible, and user-friendly flower identification system. It can support biodiversity monitoring, aid in the automation of botanical research, and enhance educational tools in botany and ecology. Additionally, the methodologies can be adapted for similar tasks in more domains, such as agricultural monitoring for crop disease identification or automated sorting in recycling operations.